"Design and Development of Novel Techniques for Clustering and Classification of Data"

Synopsis

Submitted to

Swami Ramanand Teerth Marathwada University,
NANDED

for

the registration of
Doctor of Philosophy in Computer Science
(Faculty of Computer Science & Technology)

by

Rajesh Sudhakar Walse
School of Computational Sciences
Swami Ramanand Teerth Marathwada University,
Nanded

Supervisor

Dr. G. D. Kurundkar
Assistant Professor (Computer Science)
Shri. Guru Buddhi Swami Mahavidyalaya, Purna District, Parbhani

Place of Work (Research Centre)

School of Computational Sciences
Swami Ramanand Teerth Marathwada University,
Nanded

DESIGN AND DEVELOPMENT OF NOVEL TECHNIQUES FOR CLUSTERING AND CLASSIFICATION OF DATA

Abstract:

By clustering the data, people can obtain the data distribution, observe the character of each cluster, and make further study on particular clusters. In addition, cluster analysis usually acts as the preprocessing of other data mining operations. Therefore, cluster analysis has become a very active research topic in data mining. Data mining is a new technology, developing with database and artificial intelligence. It is a processing procedure of extracting credible, novel, effective and understandable patterns from database. Cluster analysis is an important data mining technique used to find data segmentation and pattern information... As the development of data mining, a number of clustering methods have been founded, The study of clustering technique from the perspective of statistics, based on the statistical theories, our paper make effort to combine statistical method with the computer algorithm technique, and introduce the existing excellent statistical methods, including factor analysis, correspondence analysis, and functional data analysis, into data mining. The present study is undertaken to develop a Data Mining workflow using clustering and classification of data, solving clustering problem as well as extracting potentially interesting association rules. Use the appropriate proximity measure, and to select the optimal clustering model to solve clustering problems. Develop a Data Mining workflow to extract potentially interesting association rules.

1. A brief resume of proposed work / problem :

In the data-mining world, clustering and classification are two types of methods. Both these methods characterize objects into groups by one or more features. The key difference between clustering and classification is that clustering is an unsupervised learning technique used to group similar instances based on features, whereas classification is a supervised learning technique used to assign predefined tags to instances based on features. Data is business. Abundant data carries abundant problems. Data mining involves various techniques and they have associated problems carried with them along. Effective data mining becomes very vital in today and tomorrow's scenarios. This becomes possible with effective usage of clustering and classification of data to make the data more robust, meaningful and usable with least wastage.
The major problems we see with huge date are;

a) Data mining algorithms need to be more efficient and scalable to effectively extract the information from huge amount of data in databases.
b) Dealing with huge datasets that require distributed approaches.
c) Mining information from heterogeneous databases and global information systems.
d) Processing of large, complex and unstructured data into a structured format.

2. Introduction to proposed research topic

Data clustering, by definition, is an exploratory and descriptive data analysis technique, which has gained a lot of attention, e.g., in statistics, data mining, pattern recognition etc. It is an explorative way to investigate multivariate data sets that contain possibly many different data types. These data sets differ from each other's in size with respect to a number of objects and dimensions, or they contain different data types etc. Undoubtedly, the data clustering belongs to the core methods of data mining, in which one focuses on large data sets with unknown underlying structure. The intention of this report is to be an introduction into specific parts of this methodology called cluster analysis. So-called partitioning-based clustering methods are flexible methods based on iterative relocation of data points between clusters. The quality of the solutions is measured by a clustering criterion. At each iteration, the iterative relocation algorithms reduce the value of the criterion function until convergence. By changing the clustering criterion, it is possible to construct robust clustering methods that are more insensitive to erroneous and missing

Datasets Used in the Experiments for clustering algorithms:

In order to study the performance of the clustering algorithms presented in the previous Section, we applied those algorithms to a benchmark consisting of the datasets listed in Table: The first 11 datasets were obtained from University of California at Irving Machine Learning Repository [23]. The last two datasets we named Complex9 and Oval10, are two-dimensional spatial datasets whose examples distribute in different shapes. These two 2D datasets were obtained from the authors of [18] and some of them seem to be similar to proprietary datasets used in [15]. These datasets are used mainly for visualization purposes presented later in this Section.

Table: Datasets used in the benchmark

Dataset Name	No. of Examples	No. of Attributes	No. of Classes
IRIS plants	150	4	3
Glass	214	9	6
Diabetes	768	8	2
Waveform	5000	21	3
Ionosphere	351	34	2
Heart-H	294	13	2
Image Segmentation	2100	19	7
Vehicle Silhouettes	846	18	4
Vote	435	16	2
Vowel	990	13	11
Heart-StatLog	270	13	2
Complex9	3031	2	9
Oval10	3359	2	10

Clustering is a process of partitioning a set of data (or objects) into a set of meaningful sub-classes, called clusters.

Clustering has wide applications in

1. **Economic Science (especially market research).**
2. **Document classification**
3. **Pattern Recognition.**
4. **Spatial Data Analysis, create thematic maps in GIS by clustering feature spaces**
5. **Image Processing**

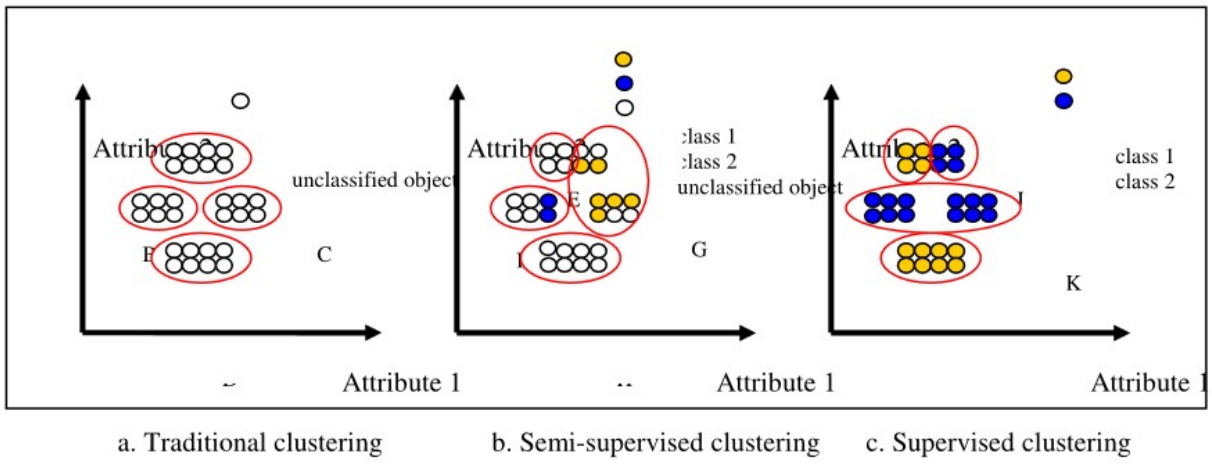
Iterative Self-Organizing Data Analysis Technique (ISODATA) method used a set of rule-of-thumb procedures that have incorporated into an iterative classification algorithm. ... The ISODATA algorithm is a modification of the k-means clustering algorithm(overcomes the disadvantages of k-means) was developed by Geoffrey H. Ball and David J. Hall, working in the Stanford Research Institute in Menlo Park, CA. They published their findings in a technical report entitled: ISODATA, a novel method of data analysis and pattern classification.

Single Representative Insertion/Deletion Steepest Decent Hill Climbing with Randomized Restart (SRIDHCR):

This greedy algorithm starts by randomly selecting a number of examples from the dataset as the initial set of representatives. Clusters are then created by assigning examples to the cluster of their closest representative. Starting from this randomly generated set of representatives, the algorithm tries to improve the quality of the clustering by adding a single non-representative example to the set of representatives as well as by removing a single representative from the set of representatives. The algorithm terminates if the solution quality (measured by q(X)) does not show any improvement. Moreover, we assume that the algorithm is run r (input parameter) times starting from a randomly generated initial set of representatives each time, reporting the best of the r solutions as its final result. The pseudo-code of algorithm SRIDHCR that was used for the evaluation of supervised clustering is given in Figure. It should be noted that the number of clusters k is not fixed for SRIDHCR; the algorithm searches for "good" values of k.

REPEAT r TIMES
curr := a randomly created set of representatives (with size between c+1 and 2c)
WHILE NOT DONE DO
1. Create new solutions S by adding a single non-representative
to curr and by removing a single representative from curr
2. Determine the element s in S for which q(s) is minimal (if
there is more than one minimal element, randomly pick one)
3. IF q(s)<q(curr) THEN curr:=s
ELSE IF q(s)=q(curr) AND |s|>|curr| THEN curr:=s
ELSE terminate and return curr as the solution for this run.
Report the best out of the r solutions found.

Table: Pseudocode for algorithm SRIDHCR



a. Traditional clustering b. Semi-supervised clustering c. Supervised clustering

Table: Traditional, Semi-Supervised, and supervised clustering

3. Survey or review of literature and development in the relevant subject:

Although the intuitive idea behind cluster analysis is simple, the successful completion of the tasks presume a large number of correct decisions and choices from several alternatives. Anderberg [5] states that there appears to be at least nine major elements in a cluster analysis study before the results can be attained. Because the current real-world data sets contain missing values as well, we complete this element list with data presentation and missing data strategy

1. Data presentation.
2. Choice of objects.
3. Choice of variables.
4. What to cluster data units or variables.
5. Normalization of variables.

6. Choice of (dis)similarity measures.
7. Choice of clustering criterion (objective function).
8. Choice of missing data strategy.
9. Algorithms and computer implementation (and their reliability, e.g., convergence)
10. Number of clusters.
11. Interpretation of results.

These are the most significant parts of the general clustering process. Jain et al. [71] suggest that the strategies used in data collection, data representation, normalization and cluster validity are as important as the clustering strategy itself. According to Hastie et al. [58, p.459], choice of the best (dis)similarity measure is even more important than the choice of clustering algorithms. This it could be also completed by validation of the resulting cluster solution [70, 4]. Validation is, on the other hand, closely related to the estimation of the number of clusters and to the result interpretation. For example, the visual exploration of the obtained solutions can be considered a kind of validation technique.
Many different method and algorithms are available:

1. For numeric and/or symbolic data
2. Exclusive vs. overlapping
3. Crisp vs. soft computing paradigms
4. Hierarchical vs. flat (non-hierarchical)
5. Access to all data or incremental learning
6. Semi-supervised mode

Algorithms also vary by:
• Measures of similarity
• Linkage methods
• Computational efficiency

Preparation of meaningful and more robust algorithms for effective clustering and classification of data, which in turn can be used for more effective data mining using computing methodologies. The list of techniques, which can be considered under such a definition, ranges from link analysis/associations, sequential patterns, analysis of time series, and classification by decision trees or neural networks, cluster analysis to scoring models. Hence, old and well-known statistical and mathematical as well as neural network methods get their new or resurrected role in Data Mining or more general in Business Intelligence. Since the 90s of the 20th century, with the information technology and the rapid development of Database technology, people can Very easy to access and store large amounts of data. The face of large-scale mass data, traditional data analysis work With only some surface treatment, but cannot get the inherent relationship between the data and the underlying information, from Fall into the "data rich, knowledge poor" dilemma [1]. To escape this dilemma, people urgently need a species can intelligently and automatically transform the data into useful information and knowledge of techniques and tools, which are on the strong force the urgent needs of data analysis tools make data mining (Data Mining) technology emerged [2]. Data mining in recent years with the database and artificial intelligence developed a new technology that the big amount of raw data to discover the hidden, useful information and knowledge to help policy makers to find the potential between the data Associated factors found to be ignored. Data mining because of its huge business prospect, are now becoming an international data library and information policy-making in the field of

cutting-edge research, and caused extensive academic and industry relations note [3]. At present, data mining has been in business management, production control, electronic commerce, market analysis and scientific science and many other fields to explore a wide range of applications [4]. The face of huge amounts of data, the first task is to sort them out, cluster analysis is to classify the raw data as a reasonable way. The so-called clustering is a group of physical or abstract objects, according to the degree of similarity between them, divided into several groups, and makes the same data objects within a group of high similarity, and different groups of data objects are not similar [5][6]. As an important function of data mining, clustering analysis can serve as a stand-alone tool to get data on the distribution of observed characteristics of each class, focus on a specific class to do some further analysis Data mining aims to discover hidden from the database, meaningful knowledge, mainly into the following categories function [12]: (1). Concept description Concept description is called as summary description, which aims to concentrate the data, given its comprehensive descriptions, or will compare it with other objects. By summing up the data, you can achieve an overall grasp of the data. Description of the most simplest concept is the use of statistics in the traditional method to calculate the various data items in the database total, mean, variance, etc., or use OL "(On Line Processing, online analytical processing) achieve multi-dimensional query and calculation of data. (2). Correlation Analysis Correlation analysis found that large amounts of data items from the set of interesting association or correlation between the contacts. With the large number of continuously collect and store data and many people in the industry from their database for mining association rules increasingly the more interesting. Records from a large number of business services found interesting correlation could help many business decisions making. (3). Classification and Prediction Classification and prediction are two forms of data analysis can be used to extract models describing important data classes or pre-future trends measured data. Classification and Prediction of a wide range of applications, for example, you can create a classification model. On the bank's loan customers to classify, to reduce the risk of the loans; also through the establishment of the classification model the functioning of the factory machines to classify, to predict the occurrence of machine failure. (4). Cluster Analysis Category according to maximum similarity and minimize between-cluster similarity principle, makes the same class of objects with high similarity with other classes of objects is very similar. each cluster formation. Class can be seen as an object class, which it can export rules. Clustering is also easy to observe the contents of the organization into hierarchical structure to organize similar events together. (5). Outlier Analysis Database may contain data objects, their general behavior with the data or the model inconsistent. This data objects are outliers. Many data mining algorithms attempt to minimize the impact of outliers, or row. In addition to them, however, in some applications may be an isolated point of a very important message. For example, in fraud detection, isolated points may indicate fraud. (6). Time Series Analysis In time series analysis, the data attribute value is changing over time. These data generally equal time intervals to obtain, but cannot get equal time intervals. Through the time, series map can be time-series data visualization. There are three basic functions in time series analysis: First, dig mode excavation, that is, by analyzing the time series of historical patterns to study the behavioral characteristics of affairs. Second, trend analysis that is, using historical data for time series forecasting the future value. Third, similarity search, which uses distance measures to ensure, given the similarity of different time series.

4. Objectives of the research and proposed hypothesis:

Keeping the above problems to be addressed in clustering algorithms presently used in data mining our main objective of the study is ***Design and development of novel techniques for clustering and classification of data for effective data mining ISO data clustering algorithms***

To achieve this objective, the following specific sub objectives have been defined

- *To identify various clustering algorithms for different purposes*
- *To classify the clustering algorithms based on linkage methods they adopt*
- *To analyses the clustering algorithms for their speed ,accuracy and computational efficiency*

Hypothesis:

Ho: There is no difference between the **data mining *ISO data clustering algorithms***

5. Methodology to be used and adopted:

Unlike other types of data, Text data have many features, in a research we have design and develop Novel techniques of feature extraction techniques using clustering method. At first, text documents like agglomerative technique, divisive techniques and distributive clustering are studied and **Novel Techniques on Feature Clustering Algorithms for text classification like** News: Electronic New articles are generated very frequently are studied and then Manual classification of these articles is a very difficult, so, automated methods are useful in this Case. This application is known as text filtering. Is studied, following this Digital Libraries are studied using a variety of supervised methods may be used for document organization in domains like digital libraries, web collections and scientific literature

It is also proposed use SPMF is an open-source data mining mining library written in Java, specialized in pattern mining. It is distributed under the GPL v3 license.
It offers implementations of 133 data mining algorithms for:

- association rule mining,
- item set mining,
- sequential pattern mining,
- sequential rule mining,
- sequence prediction,
- periodic pattern mining,
- high-utility pattern mining,
- clustering and classification,
- Time-series mining.

The following Clustering methods and algorithms are proposed to be analysed during the investigation

- the original **K-Means** algorithm (MacQueen, 1967)
- the **Bisecting K-Means** algorithm (Steinbach et al, 2000)
- algorithms for **density-based clustering**
  - the **DBScan** algorithm (Ester et al., 1996)
  - the **Optics** algorithm to extract a **cluster ordering of points,** which can then be use to generate DBScan style clusters and more (Ankerst et al, 1999)
- a **hierarchical clustering** algorithm
- a tool called **Cluster Viewer** for **visualizing clusters** New!
- a tool called **Instance Viewer** for **visualizing the input of clustering algorithms** New!

Clustering algorithms

- Example 77 : Clustering Values with the K-Means algorithm
- Example 78 : Clustering Values with the DBScan algorithm
- Example 79 : Using Optics to extract a cluster-ordering of points and DB-Scan style clusters
- Example 80 : Clustering Values with the Bisecting K-Means algorithm
- Example 81 : Clustering Values with a Hierarchical Clustering algorithm
- Example 82 : Visualizing clusters using the Cluster Viewer
- Example 83 : Visualizing clusters using the Instance Viewer

A. K- Means Method (Single Representative Insertion/Deletion Steepest Decent Hill Climbing with Randomized Restart (SRIDHCR) :

1. Select k points as the initial centroids in a random way.
2. (Re) allocate all objects to the closest centroid.
3. Recalculate the centroid of every cluster.
4. Repeat steps two and three until a termination criterion is met.
5. Pass the clarification to the next stage.

B. HAC (Hierarchical Agglomerative Clustering)

1. Calculate the proximity matrix containing the distance between every pair of patterns. Treat every pattern as a cluster.
2. Find the majority comparable pair of clusters using the proximity matrix. Combine these two clusters into one cluster.
Modernize the proximity matrix to reflect this merge operation.
3. If all patterns are in one cluster, stop. Otherwise, go to step two.

C. EM (Expectation Maximization)

Although the fact that EM can infrequently get stuck in a limited maximum as you guess the parameters by maximizing the
Log likelihood of the experimental data, in my mind there are three things that construct it magical:
» The capability to concurrently optimize a huge number of variables.
» The ability to discover good estimates for any absent information in your data at the same time.
» In the situation of clustering multidimensional data that lends itself to modeling by a Gaussian mixture, the ability to
Generate both the traditional “hard” clusters and not-so- traditional “soft” clusters.

Main Categories of Clustering Methods

• Partitioning algorithms:

Construct various partitions and then evaluate them by some criterion.

• Hierarchy algorithms: Create a hierarchical decomposition of the set of data (or objects) using some criterion.

• Density-based: based on connectivity and density functions

• Grid-based: based on a multiple-level granularity structure

• Model-based: A model is hypothesized for each of the clusters and the idea is to find the best fit of that model to each other

A good clustering method will produce high quality clusters in which:

• The intra-class (that is, intra-cluster) similarity is high. Intra

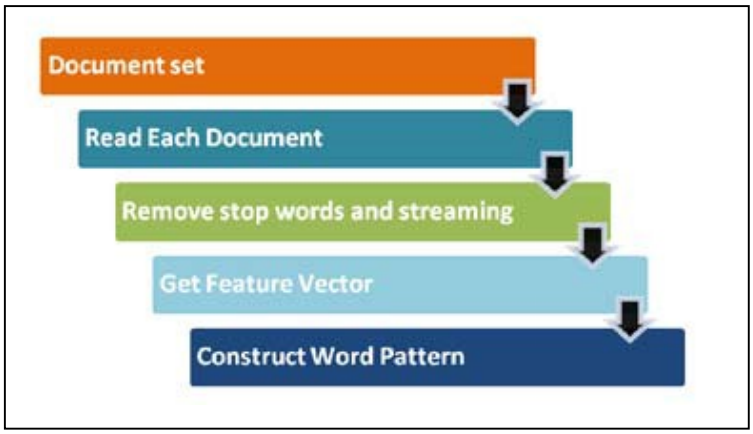
• The inter-class similarity is low.

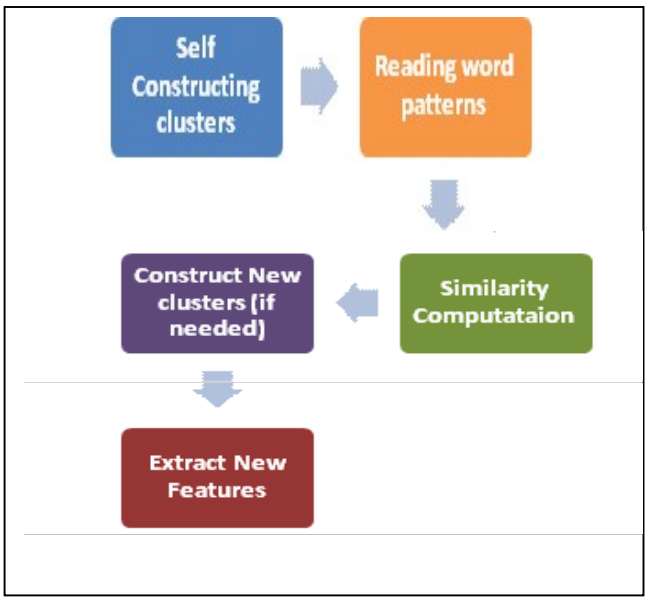
• The quality of a clustering result also depends on both the similarity measure used by the method and its implementation.

• The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns.

ANALYTICAL FRAMEWORK OF THE STUDY

The detailed process is explained below:





Process of New Feature discovery: --

Standardization / Normalization

• If the values of attributes are in different units then it is likely that some of them will take very large values, and hence the "distance" between two cases, on this variable, can be a big number.

• Other attributes may be small in values, or not vary much between cases, in which case the difference between the two cases will be small.

• The attributes with high variability / range will dominate the metric.

• Overcome this by standardization or normalization:

6. Importance of the study

Cluster analysis or Clustering is said to be a collection of objects. It is used in various application in the real world. Such as data/text mining, voice mining, image processing, web mining and so on. It is important in real world in certain fields. How and why is important in the real world and how were the techniques implemented in several applications are presented.

❖ International Status :

Customer analysis is crucial phase for companies in order to create new campaign for their existing customers. If a company can understand customer features and make efforts to fulfill their wants and provide friendly service then the customer will be supportive to the enterprise. The aim of this study was to develop a methodology to identify the characteristics of customers. It involved identification of the demographic characteristics of customers based on the analysis of categorical data using data mining clustering methods. The extracted knowledge can help companies identify valuable customers, and enable companies to make efficient knowledge-driven decisions. Customer analysis is crucial phase for companies in order to create new campaign for their existing customers. Companies are able to group or cluster certain customers, which have similar features. This may assist companies to make better marketing strategies over certain customer groups. Companies recognize that their existing customer database is their most important asset (Athanassopoulos, 2000). It is also important that how to effectively process and use customer data. Thus, this new techniques to assist analyze, comprehend or even visualize the massive amounts of stored data obtained from business and scientific applications (Liao et al, 2004). Data mining is the process of discovering and extracting considerable customer knowledge, such as rules, patterns, associations, clusters, and significant structures from large amounts of data stored in databases (Liao et al., 2008; Coussement et al., 2010). According to a research conducted by Reinartz et al., it is more beneficial to keep and satisfy existing customers than to constantly attract new customers who are characterized by a high attrition rate (Reinartz et al., 2003). Thus, if a company can understand customer features and make efforts to fulfill their wants and provide friendly service then the customer will be supportive to the enterprise. For instance, specific measures and motivation may be proposed to the most risky customer groups, i.e. the most disposed to leave the company, they may remain constant (Burez et al., 2007).

Classification and patterns extraction from customer data is very important for business support and decision-making. Timely identification of newly emerging trends is very important in business process. Large companies are having huge volume of data but starving for knowledge. To overcome the organization current issue, the new breed of technique is required that has intelligence and capability to solve the knowledge scarcity and the technique is called Data mining. The objectives of this paper are to identify the high-profit, high-value and low-risk customers by one of the data mining technique – customer clustering. In the first phase, cleansing the data and developed the patterns via demographic clustering algorithm using IBM I-Miner. In the second phase, profiling the data, develop the clusters and identify the high-value low-risk customers. This cluster typically represents the 10-20 percent of customers, which yields 80% of the revenue. Sankar Rajagopal 2011

❖ National Status :

Data analysis plays an important role in understanding various phenomena.Clustering has a significance attention in data analysis, image recognition, control process, data management, data mining etc. Due to an enormous increment in the assets of computer and communication technology. Cluster analysis aims at identifying groups of similar objects and, therefore helps to discover distribution of patterns and interesting correlations in large datasets. This review paper acts as a catalyst in the initial study of the various researchers who directly or indirectly deals with clustering in their research work. In this paper, a comprehensive study of clustering is done along with its all techniques and a simple comparison of them, so that it is easy for someone to pick a specific method as per suitable to the working environment.(Sharma.M &Kamal borane, 2014)

❖ Regional Status :

Clustering is the grouping together of similar data items into clusters. Clustering analysis is one of the main analytical methods in data mining; the method of clustering algorithm will influence the clustering results directly. This paper discusses the various types of algorithms like k-means clustering algorithms, etc…. and analyzes the advantages and shortcomings of the various algorithms. In each type, we can calculate the distance between each data object and not all cluster centers in each iteration, which makes the efficiency of clustering is high. This paper provides a broad survey of the most basic techniques and identifies .This paper deals with the issues of clustering algorithm such as time complexity and accuracy to provide the better results based on various environments. The results are discussed on huge datasets. Mythili S1, Madhiya E2 2014

The World is overflowing with various kind of data like - scientific data, environmental data, financial data, and mathematical data. Manually analyzing, classifying, and pruning of the data is a tedious task for human, because the data is growing at a faster speed in this age of network and information sharing. Clustering is important in data analysis and data mining applications. A clustering method groups the data set into several data set based on the concept of maximizing the intra- class similarity and minimizing the inter- class similarity. This paper analyze or give an overview or review about the various clustering methods: Partitioning method, hierarchical method, Density based method, Grid based method, and Model based method in data mining. Kavita Nagar 2015

❖ Scope of Study :

The purpose of clustering and classification algorithms is to make sense of and extract value from large sets of structured and unstructured data. If we are working with huge volumes of unstructured data, it only makes sense to try to partition the data into some sort of logical groupings before attempting to analyze it.

Clustering and classification allows us to take a sweeping glance of data en masse, and then form some logical structures based on what you find there before going deeper into the nuts-and-bolts analysis.

In their simplest form, clusters are sets of data points that share similar attributes, and clustering algorithms are the methods that group these data points into different clusters based on their similarities. We see clustering algorithms used for disease classification in medical science, but we also see them used for customer classification in marketing research and for environmental health risk assessment in environmental engineering.

There are different clustering methods, depending on how the dataset is to be divided. The two main types of clustering algorithms are

Hierarchical: Algorithms create separate sets of nested clusters, each in their own hierarchal level.Partitional: Algorithms create just a single set of clusters.

7. Proposed work plan-with a tentative chapter scheme :

The following activity table implies the work plan for the study:

Sr. No	Activity	Duration in weeks
1	Review of clustering algorithms for a particular datamining problem eg., Business analytics, customer relationship management etc.	**15**
2	Identification of select algorithms for analysis and integration	**48**
3	Integration of isolate cluster algorithm for the selected problem	**10**
4	Development of function prototype in JAVA for cluster Algorithm to be applied	**45**
5	Coding and Testing the prototype	**20**
6	Analysis and interpretation	**8**
7	Report writing	**10**
	Total	**156 weeks or 3 years**

8. Year wise proposed plan of work and targets to be achieved :

Objective wise	Activity	Month & Year of		Output monitor able target(s)	% to be carried out in different years		
		Start	Completion		1	2	.3.
Identify clustering algorithm	For the problem eg., Customer relationship management	October 2017	March '2018		100 %		
classify the clustering algorithms	Using linkage method they adopt	April 2018	September 2018	interface matrix design to bridge the gap	100 %		
Integrating the clustering algorithm	testing and iteration for obtaining the solution	October 2018	October 2019	Prototype for testing		50%	50 %

9. Expected outcome

This research work can improve the performance of traditional algorithms Like K-Means and presents a hybrid approach. The frequently used algorithm is K-Means, which can deal with small convex datasets preferably. It reduces the error rate and achieves accuracy. This research compares the efficiency of these clustering and classification data mining algorithms by applying them to data sets. Experiment results will generate the best classification accuracy and show the high robustness and generalization capacity compared to the other algorithms.

10. Bibliography/ References:

1. A perusal of big data classification and Hadoop technology, International Transaction of electrical and computer engineers system, 2017, Vol 4, No.1, 26-38
2. https://www.google.co.in/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0ahUKEwinsqrmvpjWAhVHtY8KHYdPAxIQFgglMAA&url=http%3A%2F%2Fwww.dtic.mil%2Fget-tr-doc%2Fpdf%3FAD%3DAD0699616&usg=AFQjCNE-Frzqvpb8-qG_iAhvSzFRKpkEog
3. http://www.cs.put.poznan.pl/jstefanowski/sed/DM-7clusteringnew.pdf
4. http://www.ijarcsms.com/docs/paper/volume2/issue12/V2I12-0095.pdf
5. M. and Heckerman, D., "An experimental comparison of several clustering and initialization method", Technical Report MSRTR-98-06, Microsoft

6. Research, Redmond, WA, February 1998.
7. Mrs. Bharati M. Ramageri, "Data Mining Techniques and Applications," Indian Journal of Computer Science and Engineering, Vol. 1 No. 4, pp.301-305, 2010.
8. Karimella Vikram and Niraj Upadhayaya, "Data Mining Tools and Techniques: a review," Computer Engineering and Intelligent Systems, Vol 2, No.8, Pp.31-39, 2011.
9. Usama Fayyad, G. Paitetsky-Shapiro, and Padhrais Smith, "knowledge discovery and data mining: Towards a unifying framework", proceedings of the International Conference on Knowledge Discovery and Data Mining, pp. 82-22, 1996
10. A Novel Method for Text and Non-Text Segmentation in Document Images International conference on Communication and Signal Processing, April 3-5, 2013, India,
11. Novel Techniques on feature clustering algorithms for text classification IJCST Vol 4, Issue Spl-4 Oct-Dec, 2013. ISSN: 2229-4333 (Print)
12. A Novel Method for text and Non-Text segmentation in document images. International Conference on Communication and signal processing, April 3-5 , 2013 978-1-4673-4866-9/13 2013 IEEE

Rajesh S. Walse
(Name of the Student)
(Research Centre):
School of Computational Sciences,
Swami Ramanand Teerth Marathwada University, Nanded

Dr. G. D. Kurundkar
(Research Supervisor/ Guide)
Assistant Professor
Department of Computer Science,
Shri. Guru Buddhi Swami Mahavidyalaya, Purna District, Parbhani